سازمان جهاد دانشگاهی صنعتی شریف

مرکز آموزش های تخصصی کاربردی

دوره تربیت دانشمند داده (Data Scientist)

www.sharif.ac

۰۲۱-۶۷۶۴۱۹۹۹

یکی از مشاغلی که به واسطه گسترش اینترنت ایجاد شده، دانشمند داده یا دیتا ساینتیست است. امروزه تقاضا برای دانشمندان داده روز به روز در حال افزایش است. شرکت‌های مختلف از شرکت‌های صنعتی و تولیدی تا بازرگانی و مالی به دنبال متخصصان داده می‌گردند. دانشمندان داده افرادی هستند که داده ها را جمع آوری، سازماندهی، تجزیه و تحلیل می‌کنند و به افراد در صنایع مختلف کمک می‌کنند تا وظایف خود را به بهترین شکل ممکن انجام دهند. دانشمند داده متخصصی است که تخصص خود را در زمینه آمار و ساختن مدل‌های یادگیری ماشین برای پیش‌بینی و پاسخ به سؤالات کلیدی کسب‌وکار به کار می‌گیرد و مانند یک تحلیلگر داده قادر به پاک سازی، تجزیه و تحلیل و تجسم داده‌ها است.

سرفصل های دوره آموزشی تربیت دانشمند داده (Data Scientist) به شرح ذیل می باشد:

تحلیل اکتشافی

- بخش اول: مفاهیم ایده ها و ساختار:
  - خاستگاه و اهمیت تحلیل اکتشافی
  - مستندسازی تحلیل
  - ساختار تحلیل داده
  - تحلیل داده نظام مند
- بخش دوم: Data Wrangling
  - Discovering
  - Structuring
  - Cleaning
  - Enriching
  - Validating
  - Publishing
- تحلیل توصیفی داده ها بر اساس نوع داده و شناسایی ارتباطات دو و چند متغیره:
  - تحلیل داده اسمی
  - تحلیل داده کمی

- خلاصه سازی داده ها
- شناسایی ارتباطات دو و چند متغیره
- Binning و ساخت متغیر جدید

- بخش سوم: آزمون فرض وA/B test
  - تعریف فرض صفر
  - تعریف فرض جایگزین
  - سطح معنی داری
  - مقدار بحرانی
  - تصمیم گیری و تفسیرنتیجه
- بخش چهارم: ارائه نتایج
  - Data Story Telling
  - Resonate
- بخش پنجم: تحلیل داده برمبنای شواهد (Evidence based) و اعتبارسنجی تحلیل داده
  - مخاطرات EDA و راهکارهای پاسخگویی به آنها
  - تکرار پذیری
  - تعمیم پذیری
- بخش ششم: پروژه نمونه

داده کاوی با پایتون

- مقدماتی درباره داده کاوی با پایتون:

- طراحی مسائل و مجموعه داده هایی جهت شناخت داده، علم داده و کاربردهای آن در صنایع بیمه، بانک، بورس، طراحی نرم افزار با استفاده از علم داده، دیجیتال مارکتینگ هوشمند با داده کاوی، سیستم پیشنهاد دهنده وب سایت، متن کاوی در وب سایت و کاربردهای آن، تصویرکاوی و سیستم پیشنهاد موسیقی.

- بررسی ویژگی های زبان پایتون و پاسخ به این سوال که چرا از پایتون استفاده میکنیم؟ بررسی امکانات و زیرساخت های زبان پایتون به همراه جزئیات پیاده سازی برخی از قسمت ها با زبان (C)
- آشنایی عمومی با کتابخانه های موجود در زبان پایتون جهت انجام عملیات داده کاوی(Scikit Learn، Tensorflow، Py Torch، Numpy، Pandas، Matplotlib و ...)
- آشنایی کلی با حوزه ی کلان داده (Big Data)، هوش نرم و چهارچوب های مورد استفاده آن به همراه کاربرد ارتباط با علم داده

- نصب و پیاده سازی محیط های عملیاتی:
  - آشنایی با ورژن های مختلف پایتون و نصب پایتون در لینوکس یا ویندوز همراه با نصب پکیج آناکوندا وآشنایی با پکیج های مهم
  - نصب و ایجاد محیط اولیه در Visual Studio Code و ایجاد یک برنامه پایتون
  - نصب و ایجاد محیط اولیه در Jupyter و ساخت یک دفترچه پایتون
- مفاهیم پایه داده ها و ریاضی و آماری:
  - داده و درک مفهوم ویژگی(Feature)، بعد (Dimension) و ماتریس(Matrix) و درک مفهوم تنسور (Tensor) و کاربرد آن در داده کاوی
  - آشنایی و کار با کتابخانه ی Numpy و Scipy برای انجام عملیات آماری
  - آنالیز مولفه اصلی(PCA) و TSNE و کاربرد آن در نمایش داده ها و کاهش ابعاد
  - بارگزاری داده ها و تعامل با داده ها با استفاده از کتابخانه ی Pandas
- نمایش داده ها:
  - آشنایی با نمودارهای مختلف(Pie، Histogram، Bar، Line، Flow و...)وکاربرد هریک ازآنها
  - نحوه ی نمایش هیستوگرام و کاربرد آن با کتابخانه ی Matplotlib
  - نمایش داده ها به صورت تعاملی در کتابخانه ی Boken
- طبقه بندی و رگرسیون و الگوریتم های مختلف آن:
  - آشنایی با نمونه داده های طبقه بندی و کاربردهای آن

- بررسی مجموعه داده های iris (تشخیص گل های زنبق از روی ویژگی ها)، MNIST (تشخیص تصاویر دست نوشته)، Boston Housing (قیمت گذاری هوشمند خانه) به عنوان نمونه های ساده و کاربردی
- معرفی روش ها و مراجع جمع آوری داده ها و استفاده از آن
  - مثال پیش بینی هوشمند هزینه و تخمین ارزش کالا
  - مثال پیش بینی وضعیت هوا و هواشناسی
  - مثال کنترل ترافیک هوشمند با استفاده از داده های شهری
  - مثال تحلیل احساسات و استقبال/عدم استقبال کاربران از محصول یک فروشگاه با استفاده از کامنت های کاربران
  - مثال پیش بینی و توصیه محصول مورد نیاز کاربر در فروشگاه اینترنتی
  - مثال پیش بینی خرید کاربر با توجه به رفتار او در فروشگاه اینترنتی
  - مثال تشخیص هوشمند حملات هکرها به سرور
  - مثال پیشبینی هوشمند جرائم شهری و پیشگیری از وقوع جرم
  - مثال پیش بینی مصرف سوخت اتومبیل
- آشنایی و پیاده سازی طبقه بندی با الگوریتم نزدیکترین همسایه (KNN) در پایتون
- آشنایی و پیاده سازی طبقه بندی با الگوریتم ماشین بردار پشتیبان (SVM) و آشنایی با انواع مختلف پیاده سازی و پارامترهای آن در پایتون
- بررسی درخت های تصمیم(Decision Trees) و پیاده سازی آنها در حل مسائل طبقه بندی در پایتون
- آشنایی و پیاده سازی طبقه بندی با الگوریتم های ترکیبی (RandomForest،AdaBoost و ...) در پایتون
- آشنایی با الگوریتم های XGBoost و CatBoost و کتابخانه های XGBoost و CatBoost
- آشنایی با معیارهای مختلف ارزیابی کیفیت طبقه بندی
  - Accuracy

- Precision
- Recall
- F1
- ROI AUC
- و ...

- خوشه بندی و الگوریتم های مختلف آن:
  - آشنایی با نمونه داده های خوشه بندی و حل مسائل کاربردی آن
  - کاربرد و آشنایی با روش های عملی خوشه بندی:
    - مثال گروه بندی مشتریان(وب سایت و فروشگاه) با روش RFM و RFM مبتنی بر زمان
    - مثال گروه بندی تصاویر دست نوشته
    - مثال گروه بندی هوشمند مطالب وب سایت بدون استفاده از ناظر
    - مثال گروه بندی حملات هکرها به یک سرور
  - آشنایی و پیاده سازی خوشه بندی با الگوریتم KMeans
  - بررسی و پیاده سازی خوشه بندی با DBSCAN
  - آشنایی با پیاده سازی DBSCAN سلسله مراتبی و کتابخانه ی HDBSCAN
  - آشنایی و پیاده سازی خوشه بندی با الگوریتم MeanShift
  - آشنایی و پیاده سازی خوشه بندی با الگوریتم سلسله مراتبی(Hierarchical Clustering)
  - آشنایی و پیاده سازی خوشه بندی با الگوریتم خوشه بندی طیفی(Spectral Clustering)
  - آشنایی با روش های ارزیابی کیفیت خوشه ها
    - Silhouette، کالینسکلی و ...

- متوازن سازی داده ها:
  - الگوریتم های OverSampling ، SMOTE، UnderSampling و... .
- آموزش کار با گوگل Colab و اجرای برنامه ها بر روی سرورهای Googlee
- آموزش کار با وب سایت Kaggle و کسب تجربه و رزومه
- کاهش ابعاد داده ها و الگوریتم های آن:
  - PCA، UMAP،TSNE، KernelPCA
- تصویرکاوی و استفاده از تکنیک های پردازش تصویر دیجیتال (HueMoments، Histogram و Haralick) در طبقه بندی و داده کاوی تصاویر

Deep Learning

قسمت اول: مقدمه

- نحوه اجرا در گوگل کولب
- پیاده سازی KNN در پایتون
- استفاده و لود تصویر در پایتون به عنوان داده
- آشنایی و پردازش داده های ارقام دست نویس فارسی
- طبقه بندی ارقام دست نویس فارسی

قسمت دوم: شبکه عصبی

- پرسپترون (یک نورون) چیست
- شبکه عصبی: استفاده از چندین نورون و لزوم تابع فعالیت
- Softmax
- تابع هزینه
- یادگیری در شبکه های عصبی: گرادیان کاهشی و پس انتشار خطا
- نرخ یادگیری
- پیاده سازی در tensorflow/Keras
- الگوریتم های گرادیان کاهشیstochastic ،: batchوmini-batch
- کد کامل پایتون
- Dropout
- نرمال‌سازی دسته‌ای(Batch norm)

قسمت سوم: شبکه های عصبی کانولوشنالی (CNN)

- مقدمه شبکه های عصبی کانولوشنالی
- چالش Imagenet
- لزوم سلسله مراتب در شبکه های عصبی
- کانولوشن و فیلترها
- ویژگی‌های مکانی فضایی و سلسله مراتب در شبکه‌های عصبی کانولوشنی (CNNs)
- padding در کانولوشن
- مفهوم Stride در کانولوشن
- کانولوشن روی عکس رنگی
- ادغام(Pooling)
- معماری Lenet-5
- محاسبه تعداد پارامترها در یک لایه کانولوشن
- پیاده سازی یک شبکه عصبی کانولوشنالی در پایتون
- ادغام میانگین سراسری (GAP)
- طبقه‌بندی باینری و چند کلاسه
- دانلود از Kaggle در محیط Google Colab
- داده‌افزایی
- لود کردن اطلاعات از هارد Tensorflow

قسمت چهارم: معماری های مهم و معروف و انتقال یادگیری

- الکس نت
- ZFNet
- VGG
- درک کانولوشن 1 در 1
- Inception
- ResNet
- مدل های از پیش آموزش دیده درkeras application
- بازشناسی اشیاء با وبکم
- انتقال یادگیری (ترنسفر لرنینگ)
- تنظیم دقیق (Fine-tuning)

قسمت پنجم: استفاده از functional api و پیاده سازی مدلهای چند ورودی / چند خروجی

- مقدمه رگرسیون
- مثال رگرسیون: تخمین قیمت خانه
- تخمین میزان مصرف سوخت ماشین
- Functional API در کراس
- تخمین قیمت خانه با ویژگی های بصری
- استفاده از دو نوع داده ورودی (ساختار یافته و بصری) در یک شبکه عصبی
- بازشناسی و تعیین محل اشیاء (localization)

قسمت ششم: طبقه‌بندی متن، استفاده از Embedding و سیستم‌های توصیه‌گر

- پیش پردازشها در متن
- Bag-of-embedding
- Ngrams
- سری های زمانی
- RNN
- LSTM
- GRU
- Transformer
- سیستم توصیه گر مبتنی برembedding

یک پله بالاتر از تخصص ...

www.sharif.ac

۰۲۱-۶۷۶۴۱۹۹۹